

PHÂN LOẠI BÀI VIẾT CỦA VOZ.VN

Môn học: Nhập môn khoa học dữ liệu

Bài làm được thực hiện bởi: Nhóm 41

- 1712495: Nguyễn Quang Huy
- 1712858: Nguyễn Ngọc Tú



GIỚI THIỆU

- Bài toán
- Hướng giải quyết

BÀI TOÁN

Phân loại bài viết cho voz.vn để chống trường hợp đặt bài viết sai forum

HƯỚNG TIẾP CẬN

- Thu thập dữ liệu: Sử dụng thư viện Scrapy
- Xây mô hình: Sử dụng deep learning Bidirectional LSTM

THU THẬP DỮ LIỆU

- Thu thập dữ liệu bài đăng từ website <https://voz.vn>
- Ở website này, người dùng sẽ tạo bài đăng vào các forum có sẵn như “Điện thoại di động”, “Phim / Nhạc / Sách”, “Thể dục thể thao”, ...
- Ta sẽ lấy về nội dung text trong bài viết của người tạo ra bài đăng, và tên của forum mà bài đăng đó thuộc về
 - > Xây dựng mô hình tự động phân loại bài viết vào forum phù hợp dựa vào nội dung

theNEXTvoz

https://voz.vn

News Forums Latests Members pik.vn F17 F33 BDVN BDQT LT Log in Register

Sản phẩm công nghệ

	Threads	Messages		
Android	1.7K	50.9K	thảo luận	nhờ các bác tư vấn ... 1 minute ago · lastfridaynight
Apple	1.1K	28K	thảo luận	Dùng tai nghe cổng... 19 minutes ago · lastfridaynight
Multimedia	575	15.1K	kiến thức	[TWS] Hội tai nghe ... Today at 8:42 PM · Sử Phú Thọ
Đồ điện tử & Thiết bị gia dụng	1.4K	39.4K	thảo luận	Official Thread về D... 28 minutes ago · Sous.levent
Chụp ảnh & Quay phim	237	14.8K	thảo luận	Hội máy ảnh Sony ... 24 minutes ago · kid27041994

Học tập & Sự nghiệp

	Threads	Messages		
Ngoại ngữ	324	10.2K	thảo luận	Cách học từ mới ch... Today at 8:29 PM · ghost_killer
Lập trình / CNTT	1.2K	29.9K	thảo luận	Không giải code thì ... 14 minutes ago · thanhdongguy
Kinh tế / Luật	326	30K	thảo luận	[CLB Chứng khoán]... 31 minutes ago · KMS_komaci
Make Money Online	635	11.3K	kiến thức	Ola City - Nền tảng ... 4 minutes ago · giangtranbang

Staff online

thanhvova
made for silence...

Trang chủ của voz.vn. Có thể thấy ví dụ một vài forum như **Apple**, **Ngoại Ngữ**,...

File Edit View History Bookmarks Tools Help

theNEXTvoz

tin tức - Đây là lý do vì sao

https://voz.vn/t/day-la-ly-do-vi-sao-cac-thiet-bi-nam-2021-cua-samsung-lai-dat-viec-tuy-chinh-len-hang-dau

VOZ

News Forums Latests Members pik.vn F17 F33 BDVN BDQT LT Log in Register

New posts

Forums > Sản phẩm công nghệ > Đồ điện tử & Thiết bị gia dụng >

tin tức Đây là lý do vì sao các thiết bị năm 2021 của Samsung lại đặt việc tùy chỉnh lên hàng đầu

lebao · Yesterday at 12:59 PM

Yesterday at 12:59 PM #1

lebao
★★★★★
Tayto
Editor

Các sự kiện của năm 2020 về cơ bản đã thay đổi ngành thiết bị gia dụng toàn cầu. Bỗng nhiên nhà của chúng ta được chuyển đổi thành văn phòng, nhà hàng, phòng tập gym và hơn thế nữa, và giờ đây chúng ta dành nhiều thời gian hơn để tương tác với các thiết bị của mình. Trong suốt năm qua, Samsung đã lắng nghe phản hồi của người dùng để hiểu những thay đổi này đã làm thay đổi kỳ vọng của họ về thiết bị gia dụng như thế nào và phân tích những thông tin đó để chọn ra được chiến lược trọng tâm cho năm 2021 chính là: sự tùy chỉnh.

JaeSeung Lee, Tổng giám đốc kiêm Trưởng bộ phận kinh doanh thiết bị gia dụng kỹ thuật số tại Samsung cho biết: “Người tiêu dùng, đặc biệt là thế hệ millennials, đang dành nhiều thời gian ở nhà hơn và yêu cầu các sản phẩm có thể nâng cao cuộc sống hàng ngày bằng các tính năng tùy chỉnh. Hiệu suất cao và đáng tin cậy hiện đã trở thành một kỳ vọng cơ bản. Tiêu chuẩn trải nghiệm gia đình tốt hơn và thông minh hơn hiện đang được quyết định bởi mức độ gắn kết giữa các thiết bị “phù hợp” với phong cách sống của khách hàng.”

Giới thiệu một Thiết kế Thiết bị Gia dụng Phù hợp hơn

Sự tùy chỉnh đã trở thành một tính năng chính cho các thiết bị của Samsung. Được dẫn dắt bằng một triết lý thiết kế được giới thiệu lần đầu tiên trong một dự án có tên Project PRISM, các thiết bị của Samsung có thể đáp ứng các nhu cầu riêng của người dùng, phản ánh phong cách sống của họ như một lăng kính khúc xạ ánh sáng thành nhiều màu sắc khác nhau.

Share this page

f t d p w s

Ví dụ một bài đăng thuộc forum **Đồ điện tử & Thiết bị gia dụng**, với tiêu đề: *Đây là lý do vì sao....hàng đầu*

THU THẬP DỮ LIỆU

Dữ liệu được thu thập bằng thư viện
Scrapy(<https://scrapy.org/>) trong Python

Scrapy hỗ trợ việc tạo http request, parse html, lưu kết quả vào file với nhiều định dạng khác nhau. Do đó, với những website có nội dung “tĩnh” như voz.vn thì không cần sử dụng thêm thư viện nào khác.

THU THẬP DỮ LIỆU

Các cột dữ liệu thu thập:

- postTitle: Tiêu đề bài viết
- postContent: Nội dung bài viết
- postLink: URL của bài viết
- forumName: tên forum bài viết thuộc về

9

Tuy nhiên, trong phần xây dựng mô hình thì chỉ sử dụng postContent và forumName

XÂY DỰNG MÔ HÌNH

Tiền xử lý dữ liệu:

Xóa các dữ liệu có cột nội dung trống

Tách từ:

Do tiếng việt không thể dùng khoảng trắng để phân tách từ, hơn nữa trong một số trường hợp tách từ phức tạp như: (Học sinh học sinh học.) => (, Học sinh, học, sinh học, .,)

=> Cần phải có mô hình tách từ hiệu quả

10

Sử dụng thư viện ViTokenizer

Link: <https://pypi.org/project/pyvi/>

Với độ chính xác 0.985

XÂY DỰNG MÔ HÌNH

- Tokenize

Sử dụng thư viện Tokenizer()

Để encode đoạn văn cần phân loại sang vector

Thư viện này sẽ tạo từ điển bằng phương thức fit.

Sau khi có từ điển thì sẽ chuyển từ sang số tương ứng trong từ điển đó.

Link: https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/text/Tokenizer

XÂY DỰNG MÔ HÌNH

- Padding vector

Vector thu được ở trên sẽ có chiều dài khác nhau, ta cần padding để các vector này có cùng độ dài.

Phương pháp padding:

- Nếu chưa đủ max len thì ta thêm 0 ở trước.
- Nếu quá max len thì ta cắt phần sau.

Thư viện hỗ trợ:

Link: https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/sequence/pad_sequences

XÂY DỰNG MÔ HÌNH

- Word Embedding

Để mô hình hiệu quả thì mỗi vector đại diện cho từ sẽ gần nhau khi biểu diễn không gian vector nếu từ tương ứng là gần nghĩa.

ví dụ học sinh, sinh viên thì sẽ gần nhau hơn so với học sinh, cầu thủ nhằm tăng độ chính xác của mô hình.

Để làm điều đó ta sử dụng pretrained model Word Embedding

Link:

<https://thiaisotaajppub.s3-ap-northeast-1.amazonaws.com/publicfiles/baomoi.window2.vn.model.bin.gz>

XÂY DỰNG MÔ HÌNH

- Xây dựng mô hình

Mô hình sử dụng gồm Embedding layer, Bidirectional LSTM, Dense (activation='softmax' => Đầu ra là xác suất cho từng class)

Layer (type)	Output Shape	Param #
embedding_3 (Embedding)	(None, 100, 300)	9052500
bidirectional_3 (Bidirection	(None, 128)	186880
dense_3 (Dense)	(None, 4)	516
Total params: 9,239,896		
Trainable params: 9,239,896		
Non-trainable params: 0		

XÂY DỰNG MÔ HÌNH

- Xây dựng mô hình

Sau khi sử dụng mô hình trình bày ở trên: Ta thu được độ chính xác là:

Training set : 99%

Test set : 90%

Có vẻ bị over fit

=> Sử dụng phương pháp chống over fit

Sử dụng hàm mất mát mới **regularized loss function**

XÂY DỰNG MÔ HÌNH

```
rnn_model.add(Dense(4, activation='softmax', kernel_regularizer=regularizers.l1_l2(l1=1e-5, l2=1e-4),  
    bias_regularizer=regularizers.l2(1e-4),  
    activity_regularizer=regularizers.l2(1e-5)))
```

Độ chính xác sau khi sử dụng chống over fit:

Training set: 99%

Test set : 92%